UCLA Statistics Preprints, Jan 2010

ON WEIGHT MATRIX AND FREE ENERGY MODELS FOR SEQUENCE MOTIF DETECTION*

By Qing Zhou†

University of California, Los Angeles

The problem of motif detection can be formulated as the construction of a discriminant function to separate sequences of a specific pattern from background. In computational biology, motif detection is used to predict DNA binding sites of a transcription factor (TF), mostly based on the weight matrix (WM) model or the Gibbs free energy (FE) model. However, despite the wide applications, theoretical analysis of these two models and their predictions is still lacking. We derive asymptotic error rates of prediction procedures based on these models under different data generation assumptions. This allows a theoretical comparison between the WM-based and the FE-based predictions in terms of asymptotic efficiency. Applications of the theoretical results are demonstrated with empirical studies on ChIP-seq data and protein binding microarray data. We find that, irrespective of underlying data generation mechanisms, the FE approach shows higher or comparable predictive power relative to the WM approach when the number of observed binding sites used for constructing a discriminant decision is not too small.

Key words: asymptotic efficiency, discriminant analysis, protein-DNA interaction, predictive error, transcription factor binding site.

1. Introduction. Transcription factors (TFs), a class of proteins, regulate gene transcription through their physical interactions with particular DNA sites. Such a DNA site is called a transcription factor binding site (TFBS), which is usually a short piece of nucleotide sequence (e.g., 'CATTGTC'). Typically, a TF can bind different sites and regulate a set of genes. A key observation is that sites of the same TF share similarity in their sequence composition, which is characterized by a motif. Since gene regulation has always been an important problem in biology, many computational methods have been developed to predict whether a given DNA sequence can be bound by a TF. Please see Elnitski et al. (2006), Ji and Wong (2006), and Vingron et al. (2009) for recent reviews on relevant methods.

The prediction of TFBS's considered in this article is formulated as a classification problem. Denote by w the width of the binding sites and code the four nucleotide bases, A, C, G and T, by a set of positive integers $\mathcal{I} = \{1, \cdots, J\}$ ($J = 4$). Suppose that we have observed a sample of labeled sequences of length w, $\boldsymbol{D}_n = \{(Y_k, \boldsymbol{X}_k)\}_{k=1}^n$, where $\boldsymbol{X}_k \in \mathcal{I}^w$ and $Y_k \in \{0, 1\}$ indicating whether $\boldsymbol{X}_k$ is bound by the TF ($Y_k = 1$) or not ($Y_k = 0$). We call $\boldsymbol{D}_n^+ = \{\boldsymbol{X}_k : Y_k = 1\}$ observed binding sites (or motif sites)

*To appear in *Journal of Computational Biology*.

†Department of Statistics, 8125 Mathematical Sciences Building, University of California, Los Angeles, CA 90095 (email: zhou@stat.ucla.edu).

and $\boldsymbol{D}_n^- = \{\boldsymbol{X}_k : Y_k = 0\}$ background sites (or background sequences). Then, motif detection is to construct a discriminant function from $\boldsymbol{D}_n$ to predict the label of any new sequence $\boldsymbol{x} \in \mathcal{I}^w$.

Most of the existing computational methods for motif detection can be classified into two groups. The starting point of the first group is the sequence specificity of binding sites, which is often summarized by the position-specific weight matrix (WM). For early developments of WM, please see Stormo (2000). Under the WM model, each nucleotide (letter) in a binding site is assumed to be generated independently from a multinomial distribution on {A, C, G, T}. This model has been widely used in search of TFBS's (e.g., Hertz and Stormo, 1999; Kel et al., 2003; Rahmann et al., 2003; Turatsinze et al., 2008), *de novo* motif finding (e.g., Stormo and Hartzell, 1989; Lawrence et al., 1993; Bailey and Elkan, 1994; Roth et al., 1998; Liu et al., 2002) and many other works reviewed in Vingron et al. (2009). The second group aims at modeling physical binding affinity between a TF and its binding sites via the concept of the Gibbs free energy (FE) or binding energy (e.g., Berg and von Hippel, 1987; Stormo and Fields, 1998; Gerland et al., 2002; Kinney et al., 2007). Assuming that each nucleotide in a DNA sequence of length w (w-mer) contributes additively to the interaction with the TF, this approach often leads to a regression-type model for the conditional distribution of binding affinity given a piece of nucleotide sequence (e.g., Djordjevic et al. 2003; Foat et al. 2006). This group of methods have tight connections with predictive modeling approaches to gene regulation, reviewed in Bussemaker et al. (2007), which can be regarded as a natural generalization to the free energy framework (Zhou and Liu, 2008). Although the standpoints are different, the two groups of approaches are in some sense closely related. They often give similar discriminant functions for predicting TFBS's, and there are many FE-based methods that use a weight matrix to approximate Gibbs free energy (e.g., Granek and Clarke, 2005; Roider et al., 2007).

In spite of the fast methodological development on the WM and the FE models, there is still a lack of solid theoretical analysis to compare model assumptions, parameter estimations and response predictions of the two approaches. Such theoretical analysis can provide insights into these methods by seeking answers to a series of questions. For example, what are the common and distinct assumptions between the WM and the FE models, what is the relative performance between the two approaches in predicting TFBS's given a certain data generation mechanism, and how to calculate their predictive error rates when the size of observed sample $\boldsymbol{D}_n$ becomes large? Without answering these questions, one may find it difficult to understand the nature of these methods and cannot extract the full information contained in extensive empirical comparisons between the two approaches.

In this article, we compare model assumptions and parameter estimations of typical WM and FE approaches, derive asymptotic error rates of their predictions under different data generation models, and perform comparative studies on large-scale binding data. The article is organized as follows. In Section 2 we review the basic models of the two approaches. Asymptotic error rates of prediction procedures based on these models are derived and analyzed in Section 3. Computational approaches are developed in Section 4 for practical applications of the theoretical results. Numerical analysis and

biological applications are presented in Sections 5 and 6, respectively, with a comparison of the WM-based and the FE-based predictions on ChIP-seq data and protein binding microarray data. The paper concludes with discussions in Section 7. Some mathematical details are provided in Appendices. Although presented in the specific context of motif detection, the results in this article are generally applicable to the modeling and classification of categorical data.

2. Models. Let c be a scalar, $\boldsymbol{u} = (u_1, \cdots, u_J)$ be a (column) vector, $\boldsymbol{v} = (v_1, \cdots, v_w) \in \mathcal{I}^w$, and $\boldsymbol{A} = (a_{ij})_{w\times J}$ and $\boldsymbol{B} = (b_{ij})_{w\times J}$ be two $w \times J$ matrices. For notational ease, we define $c \pm \boldsymbol{A} := (c \pm a_{ij})_{w\times J}$, $\boldsymbol{A}/\boldsymbol{B} := (a_{ij}/b_{ij})_{w\times J}$ provided that $b_{ij} \neq 0$, $\boldsymbol{v}\boldsymbol{A} := \sum_{i=1}^{w} a_{iv_i}$, $\boldsymbol{A}(\boldsymbol{v}) := \prod_{i=1}^{w} a_{iv_i}$ and $\boldsymbol{u}(\boldsymbol{v}) := \prod_{i=1}^{w} u_{v_i}$. Furthermore, we define $\boldsymbol{v}_{[-k]} := (v_1, \cdots, v_{k-1}, v_{k+1}, \cdots, v_w)$ and $\boldsymbol{A}_{[-k]}$ by removing the kth row from $\boldsymbol{A}$, for $k = 1, \cdots, w$. Symbols '$\xrightarrow{L}$' and '$\xrightarrow{P}$' are used for convergence in law and in probability, respectively.

Let $\boldsymbol{\theta}_0 = (\theta_{01}, \cdots, \theta_{0J})$ be the cell probabilities (probability vector) of a multinomial distribution for i.i.d. background nucleotides, where $\sum_{j=1}^{J} \theta_{0j} = 1$ and $\theta_{0j} > 0$ for $j = 1, \cdots, J$. Since $\boldsymbol{\theta}_0$ can be accurately estimated from a large number of genomic background sequences, we assume that it is given in the following analyses. Throughout the paper, we assume that the cell probabilities of any multinomial distribution are bounded away from 0.

2.1. *The weight matrix model.* Let $\boldsymbol{X} = (X_1, \cdots, X_w) \in \mathcal{I}^w$ be a sequence of length w. In the weight matrix model (WMM), we assume that $\boldsymbol{X}$ is generated from a mixture distribution. Let $Y \in \{0, 1\}$ label the mixture component. With probability q_0, $Y = 0$ and $\boldsymbol{X}$ is generated from an i.i.d. background model (with parameter) $\boldsymbol{\theta}_0$, that is, $P(\boldsymbol{X} \mid Y = 0) = \boldsymbol{\theta}_0(\boldsymbol{X})$. With probability $q_1 = 1 - q_0$, $Y = 1$ and $\boldsymbol{X}$ is generated from a weight matrix $\boldsymbol{\Theta} = (\theta_{ij})_{w\times J} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_w)^t$, where $\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{iJ})$ is a probability vector for $i = 1, \cdots, w$ and X_i is independent of other X_k ($k \neq i$). To be specific, $P(\boldsymbol{X} \mid Y = 1) = \boldsymbol{\Theta}(\boldsymbol{X})$. From this model the log-odds ratio of Y given $\boldsymbol{X}$ is

$$\log \frac{P(Y = 1 \mid \boldsymbol{X})}{P(Y = 0 \mid \boldsymbol{X})} = \log \frac{q_1 \boldsymbol{\Theta}(\boldsymbol{X})}{q_0 \boldsymbol{\theta}_0(\boldsymbol{X})} = \log(q_1/q_0) + \sum_{i=1}^{w} \log(\theta_{iX_i}/\theta_{0X_i}). \tag{1}$$

In the WM-based prediction, q_1 is typically fixed by prior expectation or determined by the relative cost of the two types of errors (false positive vs false negative). Effectively, we assume that q_1 is given. Let

$$\beta_0 = \log(q_1/q_0), \quad \beta_{ij} = \log(\theta_{ij}/\theta_{0j}), \tag{2}$$

for $1 \leq i \leq w, 1 \leq j \leq J$ and $\boldsymbol{\beta} = (\beta_{ij})_{w\times J}$. We rewrite (1) as

$$\log \frac{P(Y = 1 \mid \boldsymbol{X})}{P(Y = 0 \mid \boldsymbol{X})} = \beta_0 + \sum_{i=1}^{w} \beta_{iX_i} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} := h(\boldsymbol{X}), \tag{3}$$

which defines an additive discriminant function to predict Y given $\boldsymbol{X}$, i.e., to predict whether the sequence $\boldsymbol{X}$ can be bound by the TF. The label Y will be predicted as

1 if $h(\boldsymbol{X}) \geq 0$ and 0 otherwise. This prediction can be regarded as a naive Bayesian classifier.

Given observed binding sites $\boldsymbol{D}_n^+$, we estimate $\boldsymbol{\Theta}$ by the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\Theta}}^m = (\hat{\boldsymbol{\theta}}_1^m, \cdots, \hat{\boldsymbol{\theta}}_w^m)^t$ and substitute it in equation (2) to obtain $\hat{\boldsymbol{\beta}}^m$. Here, the superscript 'm' stands for estimators based on the WMM. Let $d\hat{\boldsymbol{\theta}}_i^m = \hat{\boldsymbol{\theta}}_i^m - \boldsymbol{\theta}_i$, which is an infinitesimal in the order of $1/\sqrt{n}$ as $n \to \infty$. The standard asymptotic theory (e.g., Ferguson 1996) implies that

$$\sqrt{n} d\hat{\boldsymbol{\theta}}_i^m \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i^m) \text{ restricted to } \sum_{j=1}^{J} d\hat{\theta}_{ij}^m = 0, \text{ as } n \to \infty, \tag{4}$$

and that $\sqrt{n} d\hat{\boldsymbol{\theta}}_i^m$, $i = 1, \cdots, w$, are mutually independent. The (j,k)th element of the covariance matrix $\boldsymbol{\Sigma}_i^m$ is $(\delta_{jk}\theta_{ij} - \theta_{ij}\theta_{ik})/q_1$ where δ_{jk} is the Kronecker delta symbol and $1 \leq j, k \leq J$. From equation (2) we have $d\hat{\beta}_{ij}^m = d\hat{\theta}_{ij}^m/\theta_{ij}$, which leads to the following limiting distribution,

$$\sqrt{n} d\hat{\beta}_{ij}^m \xrightarrow{L} \mathcal{N}(0, (1-\theta_{ij})/(\theta_{ij} q_1)), \text{ for } j = 1, \cdots, J, \text{ as } n \to \infty, \tag{5}$$

with $\sqrt{n} d\hat{\boldsymbol{\beta}}_i^m$ mutually independent for $i = 1, \cdots, w$.

2.2. *The free energy model.* Let F, $\boldsymbol{X} = (X_1, \cdots, X_w)$ and $F\boldsymbol{X}$ be a TF, a DNA sequence, and the corresponding TF-DNA complex, respectively. The process of the TF-DNA interaction can be described by the chemical reaction $F + \boldsymbol{X} = F\boldsymbol{X}$. The concentrations of the three molecules at chemical equilibrium, $[F], [\boldsymbol{X}]$ and $[F\boldsymbol{X}]$, are determined by the association constant $K_a(\boldsymbol{X})$, that is,

$$\frac{[F\boldsymbol{X}]}{[F][\boldsymbol{X}]} = K_a(\boldsymbol{X}) = \exp\left\{-\frac{\Delta G(\boldsymbol{X})}{RT}\right\},$$

where $\Delta G(\boldsymbol{X})$ is the Gibbs free energy (FE) for the interaction of F with $\boldsymbol{X}$, R is the gas constant and T the temperature. We regard $RT > 0$ as a constant. Suppose that the contribution of a single nucleotide X_i to the FE is additive (von Hippel and Berg, 1986; Benos et al., 2002) so that we may write $-\Delta G(\boldsymbol{X})/(RT) = c + \sum_{i=1}^{w} b_{iX_i}$. Then we have

$$\log \frac{[F\boldsymbol{X}]}{[\boldsymbol{X}]} = \log[F] + c + \sum_{i=1}^{w} b_{iX_i} := b_0 + \sum_{i=1}^{w} b_{iX_i}. \tag{6}$$

To avoid non-identifiability in estimation, we take $\boldsymbol{S}_{ref} = (s_1, \cdots, s_w)$ as a reference sequence to determine a baseline level of the FE, and define $\tilde{\beta}_{ij} = b_{ij} - b_{is_i}$ for all i, j and $\tilde{\beta}_0 = b_0 + \sum_{i=1}^{w} b_{is_i}$, such that $\tilde{\beta}_{is_i} \equiv 0$ for $i = 1, \cdots, w$ and

$$b_0 + \sum_{i=1}^{w} b_{iX_i} = \tilde{\beta}_0 + \sum_{i=1}^{w} \tilde{\beta}_{iX_i} \tag{7}$$

for every $\boldsymbol{X} \in \mathcal{I}^w$. Let Y be the indicator for whether $\boldsymbol{X}$ is bound by the TF at chemical equilibrium. From the physical meaning of concentration,

$$P(Y = 1 \mid \boldsymbol{X}) = \frac{[F\boldsymbol{X}]}{[\boldsymbol{X}] + [F\boldsymbol{X}]}. \tag{8}$$

Combining equations (6), (7) and (8) leads to an additive discriminant function for this free energy model (FEM),

$$\log \frac{P(Y = 1 \mid \boldsymbol{X})}{P(Y = 0 \mid \boldsymbol{X})} = \tilde{\beta}_0 + \sum_{i=1}^{w} \tilde{\beta}_{iX_i} = \tilde{\beta}_0 + \boldsymbol{X}\tilde{\boldsymbol{\beta}} := \tilde{h}(\boldsymbol{X}). \tag{9}$$

Similarly as for the WMM, we assume that $\tilde{\beta}_0$ is fixed by prior or a desired cost. Furthermore, it is conventional to assume that $\boldsymbol{X}$ is sampled from an i.i.d. background model $\boldsymbol{\theta}_0$, i.e., $P(\boldsymbol{X}) = \boldsymbol{\theta}_0(\boldsymbol{X})$. The data generation process of the FEM has a clear biological meaning. Suppose that we have sampled n nucleotide sequences of length w, $\{\boldsymbol{X}_k \in \mathcal{I}^w\}_{k=1}^n$, from the genomic background $\boldsymbol{\theta}_0$. We mix these sequences with TF molecules in a container where the concentration of the TF is held as a constant. At chemical equilibrium we label the sequences $\boldsymbol{X}_k$ bound by the TF as $Y_k = 1$ and otherwise $Y_k = 0$. The output of this experiment is the labeled sample $\boldsymbol{D}_n = \{(Y_k, \boldsymbol{X}_k)\}_{k=1}^n$. Although there exist other models based on binding free energy, we focus on this basic model in this paper, which makes a theoretical analysis relatively clean while capturing main characteristics of FE-based approaches.

Given $\boldsymbol{D}_n$, the MLE of $\tilde{\boldsymbol{\beta}}$, denoted by $\hat{\boldsymbol{\beta}}^f = \tilde{\boldsymbol{\beta}} + d\hat{\boldsymbol{\beta}}^f$ with the superscript 'f' for FE-based estimators, can be calculated by the standard logistic regression. Note that $\hat{\boldsymbol{\beta}}^f$ maximizes the conditional likelihood

$$P(Y \mid \boldsymbol{X}, \tilde{\boldsymbol{\beta}}) = \frac{\exp\{(\tilde{\beta}_0 + \boldsymbol{X}\tilde{\boldsymbol{\beta}})Y\}}{1 + \exp(\tilde{\beta}_0 + \boldsymbol{X}\tilde{\boldsymbol{\beta}})} \tag{10}$$

determined by equation (9). Similar to the results in Efron (1975), it is not difficult to demonstrate that $\hat{\boldsymbol{\beta}}^f$ is consistent for $\tilde{\boldsymbol{\beta}}$ with asymptotic normality,

$$\sqrt{n} d\hat{\boldsymbol{\beta}}^f \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^f), \text{ as } n \to \infty, \tag{11}$$

where $d\hat{\boldsymbol{\beta}}^f$ is regarded as a vector of $(J-1)w$ dimensions (recall that $\tilde{\beta}_{is_i} = \hat{\beta}^f_{is_i} \equiv 0$ for $i = 1, \cdots, w$). The asymptotic covariance matrix

$$\boldsymbol{\Sigma}^f = \left[\mathbb{E}_{\boldsymbol{\theta}_0}\left\{p_1(\boldsymbol{X})p_0(\boldsymbol{X})\boldsymbol{C}_{\boldsymbol{X}}\boldsymbol{C}_{\boldsymbol{X}}^t\right\}\right]^{-1}, \tag{12}$$

where $p_y(\boldsymbol{X}) = P(Y = y \mid \boldsymbol{X})$ for $y = 0, 1$, $\boldsymbol{C}_{\boldsymbol{X}}$ is a $(J-1)w$-dimensional column vector coding each X_i as a factor of J levels, and $\mathbb{E}_{\boldsymbol{\theta}_0}$ is taken with respect to (w.r.t.) the background model $\boldsymbol{\theta}_0$ that generates the sequence $\boldsymbol{X}$.

2.3. *Comparison.* Given $(\beta_0, \boldsymbol{\beta})$ in the WMM and the reference sequence $\boldsymbol{S}_{ref}$ in the FEM, if we let

$$\tilde{\beta}_0 = \beta_0 + \sum_{i=1}^{w} \beta_{is_i}, \quad \tilde{\beta}_{ij} = \beta_{ij} - \beta_{is_i}, \tag{13}$$

for $i = 1, \cdots, w,\ j = 1, \cdots, J$, then the two models have the same conditional distribution $[Y \mid \boldsymbol{X}]$ (3, 9) for any $\boldsymbol{X}$. To simplify notations, we shall denote the decision function (9) in the FEM by $h(\boldsymbol{X}) = \tilde{\beta}_0 + \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ hereafter. Except for this condition distribution, other model assumptions are different. The WMM assumes that the nucleotides in $\boldsymbol{X}$ are generated independently given its label Y. But this is not true for the FEM, in which the conditional probability of $\boldsymbol{X}$ given Y is

$$P(\boldsymbol{X} \mid Y, \text{FEM}) \propto P(Y \mid \boldsymbol{X}, \text{FEM}) P(\boldsymbol{X} \mid \text{FEM}) = \frac{\exp\{(\tilde{\beta}_0 + \boldsymbol{X}\tilde{\boldsymbol{\beta}})Y\}}{1 + \exp(\tilde{\beta}_0 + \boldsymbol{X}\tilde{\boldsymbol{\beta}})} \boldsymbol{\theta}_0(\boldsymbol{X}). \tag{14}$$

Since equation (14) cannot be written as a product of functions of X_i, this model implicitly allows dependence among $X_1, \cdots, X_w$. Consequently, the FEM may account for some observed nucleotide dependences within a motif such as in Bulyk et al. (2002), Barash et al. (2003), Zhou and Liu (2004), and Zhao et al. (2005) among others. On the other hand, under the FEM model the marginal distribution of $\boldsymbol{X}$ is simply the background nucleotide distribution, i.e., $P(\boldsymbol{X} \mid \text{FEM}) = \boldsymbol{\theta}_0(\boldsymbol{X})$, but the marginal distribution of $\boldsymbol{X}$ under the WMM is a mixture,

$$P(\boldsymbol{X} \mid \text{WMM}) = q_1 \boldsymbol{\Theta}(\boldsymbol{X}) + q_0 \boldsymbol{\theta}_0(\boldsymbol{X}). \tag{15}$$

The different model assumptions lead to different procedures for parameter estimation, in particular the coefficients $\boldsymbol{\beta}\,(\tilde{\boldsymbol{\beta}})$. As discussed in Sections 2.1 and 2.2, $\hat{\boldsymbol{\beta}}^m$ and $\hat{\boldsymbol{\beta}}^f$ are consistent under the WMM and under the FEM, respectively. Since $\hat{\boldsymbol{\beta}}^f$ maximizes the conditional likelihood $P(Y \mid \boldsymbol{X}, \tilde{\boldsymbol{\beta}})$ (10) which is identical between the two models, it is also consistent for $\boldsymbol{\beta}$ under the WMM up to the translation (13). However, $\hat{\boldsymbol{\beta}}^f$ is expected to be less efficient than $\hat{\boldsymbol{\beta}}^m$ in prediction if the WMM corresponds to the underlying data generation process, due to the ignorance of the information on $\boldsymbol{\Theta}$ contained in the marginal likelihood $P(\boldsymbol{X} \mid \boldsymbol{\Theta}, \text{WMM})$ (to be discussed in detail in Section 3.1). Conversely, if data are generated by the FEM, $\hat{\boldsymbol{\beta}}^m$ is biased and no longer consistent. We will analyze the bias and the resulting incremental error rate in later sections.

3. Theoretical results. For both WMM and FEM, the ideal decision function $h(\boldsymbol{x})$ is obtained with the true parameters of the respective models and the corresponding ideal error rate

$$R^* = \sum_{\boldsymbol{x}: h(\boldsymbol{x}) \geq 0} P(Y = 0, \boldsymbol{X} = \boldsymbol{x}) + \sum_{\boldsymbol{x}: h(\boldsymbol{x}) < 0} P(Y = 1, \boldsymbol{X} = \boldsymbol{x}). \tag{16}$$

Denote by

$$R^*(\boldsymbol{x}) = \min_{y \in \{0,1\}} P(Y = y \mid \boldsymbol{X} = \boldsymbol{x})$$

the ideal error rate for $h(\boldsymbol{x})$ given $\boldsymbol{X} = \boldsymbol{x}$. Consider a decision function $\hat{h}(\boldsymbol{x})$ estimated from $\boldsymbol{D}_n$. Given any $\boldsymbol{x}$ for which $h(\boldsymbol{x})\hat{h}(\boldsymbol{x}) < 0$, the incremental error rate beyond $R^*(\boldsymbol{x})$ is

$$\Delta R(\boldsymbol{x}) = |P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) - P(Y = 0 \mid \boldsymbol{X} = \boldsymbol{x})|\,.$$

Then the expectation of the total incremental error rate for $\hat{h}$ is

$$\begin{aligned}\mathbb{E}[\Delta R(\hat{h})] &= \mathbb{E}\left[\sum_{\boldsymbol{x}\in\mathcal{I}^w} \Delta R(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x})\, \mathbf{1}\left\{h(\boldsymbol{x})\hat{h}(\boldsymbol{x}) < 0\right\}\right] \\ &= \sum_{\boldsymbol{x}\in\mathcal{I}^w} \Delta R(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}) P\left\{h(\boldsymbol{x})\hat{h}(\boldsymbol{x}) < 0\right\}, \qquad (17)\end{aligned}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Please note that $\hat{h}$, constructed from a sample of size n, is a random function. Let $\Delta\hat{h}(\boldsymbol{x}) = \hat{h}(\boldsymbol{x}) - h(\boldsymbol{x})$ be the deviation of $\hat{h}(\boldsymbol{x})$ from $h(\boldsymbol{x})$. In what follows, we will derive two theorems on $\mathbb{E}[\Delta R(\hat{h})]$ under different assumptions for $\Delta\hat{h}(\boldsymbol{x})$. As we will see, the asymptotic error rates of the WM and the FE procedures under the data generation models discussed in this paper can all be calculated based on the two theorems.

Suppose that, for every $\boldsymbol{x}$, $\sqrt{n}\Delta\hat{h}(\boldsymbol{x}) \xrightarrow{L} \mathcal{N}(0, V(\hat{h}, \boldsymbol{x}))$, where $V(\hat{h}, \boldsymbol{x})$ is the asymptotic variance. As $n \to \infty$,

$$\begin{aligned}\mathbb{E}[\Delta R(\hat{h})] &\to \sum_{\boldsymbol{x}\in\mathcal{I}^w} \Delta R(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}) P(\Delta\hat{h}(\boldsymbol{x}) > |h(\boldsymbol{x})|) \\ &= \sum_{\boldsymbol{x}\in\mathcal{I}^w} \Delta R(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x})\, \Phi\left\{-\sqrt{nh^2(\boldsymbol{x})/V(\hat{h}, \boldsymbol{x})}\right\}, \qquad (18)\end{aligned}$$

where Φ is the cdf of the standard normal distribution $\mathcal{N}(0, 1)$. Let

$$\alpha(\hat{h}) = \min_{h(\boldsymbol{x})\neq 0} h^2(\boldsymbol{x})/V(\hat{h}, \boldsymbol{x}), \qquad (19)$$

and $\boldsymbol{x}^*$ be the corresponding minimum. Note that $\Delta R(\boldsymbol{x}) = 0$ when $h(\boldsymbol{x}) = 0$. Thus, as $n \to \infty$, $\mathbb{E}[\Delta R(\hat{h})]$ is dominated by the term $\Delta R(\boldsymbol{x}^*) P(\boldsymbol{X} = \boldsymbol{x}^*)\, \Phi\left[-\{n\alpha(\hat{h})\}^{1/2}\right]$, where $\alpha(\hat{h})$ determines the rate of convergence. Using the theory of large deviations, we obtain:

THEOREM 1. *If $\sqrt{n}\Delta\hat{h}(\boldsymbol{x}) \xrightarrow{L} \mathcal{N}(0, V(\hat{h}, \boldsymbol{x}))$ for every $\boldsymbol{x} \in \mathcal{I}^w$ then*

$$\frac{1}{n}\log \mathbb{E}[\Delta R(\hat{h})] \to -\frac{\alpha(\hat{h})}{2}, \text{ as } n \to \infty.$$

Let $\hat{h}^a$ and $\hat{h}^b$ be two estimated decisions constructed from samples of size n_a and n_b, respectively. Suppose that both of them satisfy the condition in Theorem 1. We define the *asymptotic relative efficiency* (ARE) of $\hat{h}^a$ with respect to $\hat{h}^b$ by $\mathrm{ARE}(\hat{h}^a, \hat{h}^b) = \alpha(\hat{h}^a)/\alpha(\hat{h}^b)$, which is the limit ratio n_b/n_a required to achieve the same asymptotic performance.

If $\hat{h}(\boldsymbol{x})$ is biased in the sense that $\sqrt{n}\{\Delta\hat{h}(\boldsymbol{x}) - \mu(\hat{h}, \boldsymbol{x})\} \xrightarrow{L} \mathcal{N}(0, V(\hat{h}, \boldsymbol{x}))$, where $\mu(\hat{h}, \boldsymbol{x})$ denotes the asymptotic bias of $\hat{h}(\boldsymbol{x})$, then simple derivation from equation (17) gives that

$$\mathbb{E}[\Delta R(\hat{h})] \to \sum_{\boldsymbol{x}\in\mathcal{I}^w} \Delta R(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x})\ \Phi\left[\frac{-\sqrt{n}\ \mathrm{sign}\{h(\boldsymbol{x})\}\{h(\boldsymbol{x}) + \mu(\hat{h}, \boldsymbol{x})\}}{\sqrt{V(\hat{h}, \boldsymbol{x})}}\right],$$

as $n \to \infty$, where $\mathrm{sign}(y)$ is the sign of y with $\mathrm{sign}(0) \equiv 0$.

THEOREM 2. *Suppose that* $\sqrt{n}\{\Delta\hat{h}(\boldsymbol{x}) - \mu(\hat{h}, \boldsymbol{x})\} \xrightarrow{L} \mathcal{N}(0, V(\hat{h}, \boldsymbol{x}))$ *for every* $\boldsymbol{x} \in \mathcal{I}^w$. *Let* $\mathcal{B}(\hat{h}) = \{\boldsymbol{x} : \mathrm{sign}\{h(\boldsymbol{x})\}\{h(\boldsymbol{x}) + \mu(\hat{h}, \boldsymbol{x})\} < 0\}$. *Then*

$$\mathbb{E}[\Delta R(\hat{h})] \to \sum_{\boldsymbol{x}\in\mathcal{B}(\hat{h})} \Delta R(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}),\ \ \textit{as } n \to \infty. \tag{20}$$

We ignore the case $\{\boldsymbol{x} : h(\boldsymbol{x}) + \mu(\hat{h}, \boldsymbol{x}) = 0\}$ which practically never happens. The set $\mathcal{B}(\hat{h})$ is the collection of $\boldsymbol{x}$ for which the estimated decision $\hat{h}$ gives a different predicted label from the ideal decision h as $n \to \infty$. Note that $\mathbb{E}[\Delta R(\hat{h})]$ does not vanish if $\mathcal{B}(\hat{h})$ is nonempty. Thus, the incremental percentage over the ideal error rate, $\mathbb{E}[\Delta R(\hat{h})]/R^*$, is an appropriate measure of the predictive performance of $\hat{h}$.

In the remainder of this section, we derive and compare the error rates of the WM and the FE procedures. From Sections 3.1 to 3.4, we assume that the constant term $\beta_0(\tilde{\beta}_0)$ is fixed to its true value. The results are generalized to situations where the constant is mis-specified in Section 3.5. The computation of $\alpha(\hat{h})$ (19) and $\mathbb{E}[\Delta R(\hat{h})]$ (20) will be discussed in Section 4.

3.1. *Error rates under WMM.* In this subsection we assume that the underlying data generation process is given by the WMM. Since both $\hat{\boldsymbol{\beta}}^m$ and $\hat{\boldsymbol{\beta}}^f$ are consistent with asymptotic normality under the WMM, we may uniformly denote their decision functions by $\hat{h}(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}\hat{\boldsymbol{\beta}} = h(\boldsymbol{x}) + \boldsymbol{x}d\hat{\boldsymbol{\beta}}$, where $d\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and $\sqrt{n}d\hat{\boldsymbol{\beta}}$ follows a normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$. This implies that $\sqrt{n}\Delta\hat{h}(\boldsymbol{x}) = \sqrt{n}\boldsymbol{x}d\hat{\boldsymbol{\beta}} \xrightarrow{L} \mathcal{N}(0, V^m(\hat{\boldsymbol{\beta}}, \boldsymbol{x}))$ with $V^m(\hat{\boldsymbol{\beta}}, \boldsymbol{x})$ being the asymptotic variance. The superscript 'm' indicates the WMM as the data generation model. Let $\mathbb{E}[\Delta R^m(\hat{\boldsymbol{\beta}})]$ be the expected incremental error rate of $\hat{h}$ indexed by $\hat{\boldsymbol{\beta}}$. Following Theorem 1,

$$\frac{1}{n}\log\mathbb{E}[\Delta R^m(\hat{\boldsymbol{\beta}})] \to -\frac{\alpha^m(\hat{\boldsymbol{\beta}})}{2} = -\frac{1}{2}\min_{h(\boldsymbol{x})\neq 0}\frac{h^2(\boldsymbol{x})}{V^m(\hat{\boldsymbol{\beta}}, \boldsymbol{x})}, \tag{21}$$

as $n \to \infty$. Consequently, the ARE of the FE procedure w.r.t the WM procedure, $\mathrm{ARE}^m(\hat{\boldsymbol{\beta}}^f, \hat{\boldsymbol{\beta}}^m)$, is determined by the ratio of $\alpha^m(\hat{\boldsymbol{\beta}}^f)$ over $\alpha^m(\hat{\boldsymbol{\beta}}^m)$.

The decision function of the WM procedure is constructed with $\hat{\boldsymbol{\beta}}^m$ (Section 2.1). Note that $\boldsymbol{x}d\hat{\boldsymbol{\beta}} = \sum_i d\hat{\beta}_{ix_i}$ is a summation of w $d\hat{\beta}_{ij}$'s, each from a different $d\hat{\boldsymbol{\beta}}_i$. The

limiting distribution of $\sqrt{n}d\hat{\beta}_{ij}^m$ (5) and the mutual independence among $d\hat{\boldsymbol{\beta}}_i^m$ imply that the asymptotic variance of $\sqrt{n}\boldsymbol{x}d\hat{\boldsymbol{\beta}}^m$ is

$$\frac{1}{q_1}\sum_{i=1}^{w}(1-\theta_{ix_i})/\theta_{ix_i} = \frac{1}{q_1}\boldsymbol{x}\left\{(1-\boldsymbol{\Theta})/\boldsymbol{\Theta}\right\},$$

and consequently,

$$\alpha^m(\hat{\boldsymbol{\beta}}^m) = \min_{\boldsymbol{x}\boldsymbol{\beta}\neq-\beta_0}\frac{q_1(\beta_0+\boldsymbol{x}\boldsymbol{\beta})^2}{\boldsymbol{x}\{(1-\boldsymbol{\Theta})/\boldsymbol{\Theta}\}}.$$

Suppose that we have chosen $(s_1,\cdots,s_w)$ as the reference sequence in the FE procedure. Define $\tilde{\beta}_0$ and $\tilde{\boldsymbol{\beta}}$ from the parameters $(\beta_0,\boldsymbol{\beta})$ of the WMM by equation (13). Then the FE-based estimator $\hat{\boldsymbol{\beta}}^f$ is consistent for $\tilde{\boldsymbol{\beta}}$ with asymptotic normality. Let $d\hat{\boldsymbol{\beta}}^f = \hat{\boldsymbol{\beta}}^f - \tilde{\boldsymbol{\beta}}$. Similar to equation (12), the asymptotic covariance matrix of $\sqrt{n}d\hat{\boldsymbol{\beta}}^f$ is $\left[\mathbb{E}\left\{p_1(\boldsymbol{X})p_0(\boldsymbol{X})\boldsymbol{C}_{\boldsymbol{X}}\boldsymbol{C}_{\boldsymbol{X}}^t\right\}\right]^{-1}$, where the expectation is taken w.r.t. the marginal distribution of $\boldsymbol{X}$ under the WMM (15). Thus the covariance matrix can be written as

$$\mathrm{Cov}^m\left(\sqrt{n}d\hat{\boldsymbol{\beta}}^f\right) = \left[q_0\mathbb{E}_{\boldsymbol{\theta}_0}\left\{\frac{e^{h(\boldsymbol{X})}}{e^{h(\boldsymbol{X})}+1}\boldsymbol{C}_{\boldsymbol{X}}\boldsymbol{C}_{\boldsymbol{X}}^t\right\}\right]^{-1}, \tag{22}$$

where the expectation $\mathbb{E}_{\boldsymbol{\theta}_0}$ averages over $\boldsymbol{X}\in\mathcal{I}^w$ generated from the background model $\boldsymbol{\theta}_0$. Based on equation (22), one can calculate the variance of $\sqrt{n}\boldsymbol{x}d\hat{\boldsymbol{\beta}}^f$ for every $\boldsymbol{x}$ and determine the convergence rate $\alpha^m(\hat{\boldsymbol{\beta}}^f)$ of the expected incremental error rate $\mathbb{E}[\Delta R^m(\hat{\boldsymbol{\beta}}^f)]$ for the FE procedure.

Because the estimation of $\hat{\boldsymbol{\beta}}^f$ is only based on the conditional distribution $[Y\mid\boldsymbol{X}]$ while $\hat{\boldsymbol{\beta}}^m$ is estimated from the joint distribution of Y and $\boldsymbol{X}$, we expect $\hat{\boldsymbol{\beta}}^f$ to be less efficient in prediction with $\alpha^m(\hat{\boldsymbol{\beta}}^f) < \alpha^m(\hat{\boldsymbol{\beta}}^m)$. We will conduct a numerical study in Section 5 to evaluate $\mathrm{ARE}^m(\hat{\boldsymbol{\beta}}^f,\hat{\boldsymbol{\beta}}^m)$ on 200 transcription factors to confirm our conclusion. Here we demonstrate the lower efficiency of $\hat{\boldsymbol{\beta}}^f$ by the loss of Fisher information in estimating an individual θ_{ij} from the conditional likelihood only. For simplicity, suppose that $\boldsymbol{\Theta}_{[-i]}$ is given and collapse X_i into two categories, $X_i=j$ and $X_i\neq j$. Because

$$P(\boldsymbol{X},Y\mid\boldsymbol{\Theta}) = P(Y\mid\boldsymbol{X},\boldsymbol{\Theta})P(\boldsymbol{X}\mid\boldsymbol{\Theta})$$

under the WMM, the loss of information equals the Fisher information on θ_{ij} contained in the marginal likelihood $P(\boldsymbol{X}\mid\boldsymbol{\Theta})$, denoted by $I(\theta_{ij}\mid\boldsymbol{X})$. Let $I(\theta_{ij}\mid\boldsymbol{X},Y)$ be the Fisher information on θ_{ij} given $\boldsymbol{X}$ and Y jointly. We define

$$\Delta(\theta_{ij}\mid[Y\mid\boldsymbol{X}]) = I(\theta_{ij}\mid\boldsymbol{X})/I(\theta_{ij}\mid\boldsymbol{X},Y) \tag{23}$$

as the fraction of the loss of information on θ_{ij} in the conditional likelihood $P(Y\mid\boldsymbol{X},\boldsymbol{\Theta})$.

PROPOSITION 3. *Let $\bar{\theta}_{ij} = q_0\theta_{0j}+q_1\theta_{ij}$. If $(Y,\boldsymbol{X})$ is drawn from the WMM then*

$$\Delta(\theta_{ij}\mid[Y\mid\boldsymbol{X}]) \geq \frac{q_1\theta_{ij}(1-\theta_{ij})}{\bar{\theta}_{ij}(1-\bar{\theta}_{ij})} := B(q_1,\theta_{ij},\theta_{0j}).$$

A proof of this proposition is given in Appendix A. If one chooses to include an equal number of background sites ($Y = 0$) and binding sites ($Y = 1$) in logistic regression to estimate $\hat{\boldsymbol{\beta}}^f$, which effectively specifies $q_0 = q_1 = 0.5$ by design, then this lower bound may be substantial. For example, with a uniform background distribution $\theta_{0j} = 0.25$ for $j = 1, \cdots, 4$, the range of $B(q_1, \theta_{ij}, \theta_{0j})$ is between 20% and 55% for most typical values of θ_{ij} (Table 1).

TABLE 1
Typical values of $B(0.5, \theta_{ij}, 0.25)$

θ_{ij}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$B(0.5, \theta_{ij}, 0.25)(\%)$	31	46	53	55	53	49	42	32	18

3.2. *WMM with Markov background.* We generalize the background model to a Markov chain, which often represents a better fit to genomic background in high organisms. We assume that given $Y = 0$, $\boldsymbol{X}$ is generated by a first order Markov chain with a transition probability matrix $\boldsymbol{\psi}_0 = (\psi_0(x, y))_{J \times J}$ where $x, y \in \mathcal{I}$. For any $\boldsymbol{x} = (x_1, \cdots, x_w) \in \mathcal{I}^w$, $\boldsymbol{\psi}_0(\boldsymbol{x}) := \prod_{i=1}^w \psi_0(x_{i-1}, x_i)$, where $\psi_0(x_0, x_1)$ is interpreted as the probability of x_1 under the stationary distribution of the Markov chain. The ideal decision function under this model is

$$h_1(\boldsymbol{x}) = \log \frac{P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})}{P(Y = 0 \mid \boldsymbol{X} = \boldsymbol{x})} = \beta_0 + \sum_{i=1}^{w} \log \theta_{ix_i} - \log \psi_0(x_{i-1}, x_i), \tag{24}$$

where $\beta_0 = \log(q_1/q_0)$ and the subscript '1' [in $h_1(\boldsymbol{x})$ and ΔR_1^m (26)] indicates a quantity whose definition involves a Markov background model. Since $\boldsymbol{\psi}_0$ can be accurately estimated with sufficient genomic background sequences, we assume that it is given. With the MLE $\hat{\boldsymbol{\Theta}}^m$ from observed binding sites, the WM procedure constructs a decision whose expected incremental error rate converges to zero exponentially fast as $n \to \infty$, following Theorem 1.

With a slight abuse of notations, let us denote by $\boldsymbol{\theta}_0$ the probability vector of the stationary distribution of the Markov chain, which is also the marginal distribution of any nucleotide X_i in a background site. We still define β_0 and $\boldsymbol{\beta}$ by equation (2) with $\boldsymbol{\theta}_0$ being the stationary probabilities, and translate β_0 and $\boldsymbol{\beta}$ via a reference sequence to $\tilde{\beta}_0$ and $\tilde{\boldsymbol{\beta}}$ (13). Let $(Y, \boldsymbol{X})$ be a sample from the WMM with Markov background. If the dependence among neighboring nucleotides in a background site is ignored, the conditional likelihood $P(Y \mid \boldsymbol{X}, \tilde{\boldsymbol{\beta}})$, parameterized by $\tilde{\boldsymbol{\beta}}$, is then given by the same expression in equation (10). Because the FE-based estimator $\hat{\boldsymbol{\beta}}^f$ maximizes this conditional likelihood, it is standard to show that $\hat{\boldsymbol{\beta}}^f \xrightarrow{P} \tilde{\boldsymbol{\beta}}$ and is asymptotically normal. Let $\hat{h}^f(\boldsymbol{x}) = \tilde{\beta}_0 + \boldsymbol{x}\hat{\boldsymbol{\beta}}^f$ denote the estimated decision function of the FE procedure. As $n \to \infty$,

$$\hat{h}^f(\boldsymbol{x}) \xrightarrow{P} \tilde{\beta}_0 + \boldsymbol{x}\tilde{\boldsymbol{\beta}} = \beta_0 + \sum_{i=1}^{w} \log \theta_{ix_i} - \log \theta_{0x_i}\,. \tag{25}$$

Let $\Delta\hat{h}^f(\boldsymbol{x}) = \hat{h}^f(\boldsymbol{x}) - h_1(\boldsymbol{x})$ be the deviation of $\hat{h}^f(\boldsymbol{x})$ from the ideal decision (24). Comparing equations (24) and (25) gives the asymptotic bias,

$$b(\boldsymbol{x}) = \sum_{i=1}^{w} \log \psi_0(x_{i-1}, x_i) - \log \theta_{0x_i} = \log \boldsymbol{\psi}_0(\boldsymbol{x}) - \log \boldsymbol{\theta}_0(\boldsymbol{x}).$$

Due to the asymptotic normality of $\hat{\boldsymbol{\beta}}^f$, we have $\sqrt{n}\{\Delta\hat{h}^f(\boldsymbol{x}) - b(\boldsymbol{x})\} \xrightarrow{L} \mathcal{N}(0, V_1^m(\hat{\boldsymbol{\beta}}^f, \boldsymbol{x}))$, where $V_1^m(\hat{\boldsymbol{\beta}}^f, \boldsymbol{x})$ is the corresponding asymptotic variance. Under this model,

$$\Delta R(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}) = |q_1 \boldsymbol{\Theta}(\boldsymbol{x}) - q_0 \boldsymbol{\psi}_0(\boldsymbol{x})| = q_0 \boldsymbol{\psi}_0(\boldsymbol{x}) \left| e^{h_1(\boldsymbol{x})} - 1 \right|.$$

Following Theorem 2 with $\mu(\hat{h}^f, \boldsymbol{x}) = b(\boldsymbol{x})$, the expected incremental error rate

$$\mathbb{E}[\Delta R_1^m(\hat{\boldsymbol{\beta}}^f)] \to \sum_{\mathcal{B}_1^m(\hat{\boldsymbol{\beta}}^f)} q_0 \left| e^{h_1(\boldsymbol{x})} - 1 \right| \boldsymbol{\psi}_0(\boldsymbol{x}), \text{ as } n \to \infty, \tag{26}$$

where $\mathcal{B}_1^m(\hat{\boldsymbol{\beta}}^f) = \{\boldsymbol{x} : \text{sign}\{h_1(\boldsymbol{x})\}\{h_1(\boldsymbol{x}) + b(\boldsymbol{x})\} < 0\}$. The incremental percentage over the ideal error rate, $\mathbb{E}[\Delta R_1^m(\hat{\boldsymbol{\beta}}^f)]/(R_1^m)^*$, is appropriate for comparing the FE-based prediction with the WM-based prediction whose expected error rate converges to $(R_1^m)^*$. A general expression for R^* is given in equation (16) which, under the WMM with Markov background, is written as

$$(R_1^m)^* = q_0 \mathbb{E}_{\boldsymbol{\psi}_0} \left\{ \mathbf{1}(h_1(\boldsymbol{X}) \geq 0) + e^{h_1(\boldsymbol{X})} \mathbf{1}(h_1(\boldsymbol{X}) < 0) \right\}. \tag{27}$$

3.3. *Error rates under FEM.* We now analyze asymptotic error rates of the two procedures regarding the FEM as the underlying data generation mechanism.

The FE-based estimator $\hat{\boldsymbol{\beta}}^f$ is consistent for $\tilde{\boldsymbol{\beta}}$ under the FEM. The asymptotic normality of $\sqrt{n}d\hat{\boldsymbol{\beta}}^f$ (11, 12) implies that $\sqrt{n}\boldsymbol{x}d\hat{\boldsymbol{\beta}}^f \xrightarrow{L} \mathcal{N}(0, V^f(\hat{\boldsymbol{\beta}}^f, \boldsymbol{x}))$. Let $\mathbb{E}[\Delta R^f(\hat{\boldsymbol{\beta}}^f)]$ be the expected incremental error rate of the FE procedure under the FEM. From Theorem 1, we have

$$\frac{1}{n} \log \mathbb{E}[\Delta R^f(\hat{\boldsymbol{\beta}}^f)] \to -\frac{\alpha^f(\hat{\boldsymbol{\beta}}^f)}{2} = -\frac{1}{2} \min_{h(\boldsymbol{x}) \neq 0} \frac{h^2(\boldsymbol{x})}{V^f(\hat{\boldsymbol{\beta}}^f, \boldsymbol{x})}, \text{ as } n \to \infty. \tag{28}$$

Denote by $\boldsymbol{\theta}_i^f = (\theta_{i1}^f, \cdots, \theta_{iJ}^f)$ the probability vector of the conditional distribution $[X_i \mid Y = 1]$ under the FEM, i.e.,

$$\theta_{ij}^f = P(X_i = j \mid Y = 1, \text{FEM}), \tag{29}$$

for $i = 1, \cdots, w$, and call $\boldsymbol{\Theta}^f = (\theta_{ij}^f)_{w \times J}$ the weight matrix. Recall that the WM-based estimator $\hat{\boldsymbol{\beta}}^m$ is obtained by estimating $\boldsymbol{\theta}_i^f$ individually from observed binding sites $\boldsymbol{D}_n^+$ and then transforming the estimates via equation (2). Denote the estimated weight matrix by $\hat{\boldsymbol{\Theta}}^m$. Since the data are generated by the FEM, $\hat{\boldsymbol{\Theta}}^m \xrightarrow{P} \boldsymbol{\Theta}^f$ and $\sqrt{n}d\hat{\boldsymbol{\Theta}}^m = \sqrt{n}(\hat{\boldsymbol{\Theta}}^m - \boldsymbol{\Theta}^f)$ follows a multivariate normal distribution asymptotically,

similar to (4), but $d\hat{\boldsymbol{\theta}}_i^m$ and $d\hat{\boldsymbol{\theta}}_k^m$ may be correlated ($1 \le i \ne k \le w$). Given that the coefficients $\tilde{\boldsymbol{\beta}}$ in the FEM are defined w.r.t. a reference sequence, we transform $\hat{\boldsymbol{\Theta}}^m$ to

$$\hat{\beta}_{ij}^m = \log(\hat{\theta}_{ij}^m/\theta_{0j}) - \log(\hat{\theta}_{is_i}^m/\theta_{0s_i}), \text{ for all } i, j, \tag{30}$$

where s_i is the ith nucleotide of the reference sequence $\boldsymbol{S}_{ref}$. Let $\Delta\hat{\boldsymbol{\beta}}^m = \hat{\boldsymbol{\beta}}^m - \tilde{\boldsymbol{\beta}}$ be the deviation of $\hat{\boldsymbol{\beta}}^m = (\hat{\beta}_{ij}^m)_{w\times J}$. To obtain its asymptotic distribution, we determine the cell probability θ_{ij}^f (29) from equation (14), that is,

$$\theta_{ij}^f \propto \theta_{0j} e^{\tilde{\beta}_{ij}} \sum_{\boldsymbol{x}\in\mathcal{I}^{w-1}} \frac{e^{\tilde{\beta}_0 + \boldsymbol{x}\tilde{\boldsymbol{\beta}}_{[-i]}}}{1 + e^{\tilde{\beta}_{ij}} e^{\tilde{\beta}_0 + \boldsymbol{x}\tilde{\boldsymbol{\beta}}_{[-i]}}} \boldsymbol{\theta}_0(\boldsymbol{x}) = \theta_{0j} e^{\tilde{\beta}_{ij}} \mathbb{E}_{\boldsymbol{\theta}_0}\left\{\frac{e^{U_i}}{1 + e^{\tilde{\beta}_{ij}} e^{U_i}}\right\},$$

where $U_i = \tilde{\beta}_0 + \boldsymbol{X}\tilde{\boldsymbol{\beta}}_{[-i]}$ for $\boldsymbol{X} \in \mathcal{I}^{w-1}$. In particular, $\theta_{is_i}^f \propto \theta_{0s_i}\mathbb{E}_{\boldsymbol{\theta}_0}\left\{e^{U_i}/(1+e^{U_i})\right\}$ since $\tilde{\beta}_{is_i} = 0$. For $i = 1, \cdots, w$ and $j = 1, \cdots, J$, we define

$$\delta_{ij} = \log \mathbb{E}_{\boldsymbol{\theta}_0}\left\{\frac{e^{U_i}}{1 + e^{\tilde{\beta}_{ij}} e^{U_i}}\right\} - \log \mathbb{E}_{\boldsymbol{\theta}_0}\left\{\frac{e^{U_i}}{1 + e^{U_i}}\right\} \tag{31}$$

and rewrite $\theta_{ij}^f = \theta_{0j} e^{\tilde{\beta}_{ij} + \delta_{ij}}/Z_i$, where $Z_i = \sum_j \theta_{0j} e^{\tilde{\beta}_{ij}+\delta_{ij}}$ is the normalizing constant. Because $\hat{\theta}_{ij}^m \xrightarrow{P} \theta_{ij}^f$ and $\tilde{\beta}_{is_i} = \delta_{is_i} = 0$, from equation (30) we have

$$\hat{\beta}_{ij}^m \xrightarrow{P} \log(\theta_{ij}^f/\theta_{0j}) - \log(\theta_{is_i}^f/\theta_{0s_i}) = \tilde{\beta}_{ij} + \delta_{ij} \tag{32}$$

for all i and j as $n \to \infty$. Thus, $\boldsymbol{\delta} = (\delta_{ij})_{w\times J}$ is the asymptotic bias of $\hat{\boldsymbol{\beta}}^m$. From the asymptotic normality of $\sqrt{n}d\hat{\boldsymbol{\Theta}}^m$, we see that $\sqrt{n}(\Delta\hat{\boldsymbol{\beta}}^m - \boldsymbol{\delta})$ follows a multivariate normal distribution with mean $\mathbf{0}$ and finite covariance matrix as $n \to \infty$. Note that this multivariate normal distribution is defined on a $(J-1)w$-dimensional space, since $\Delta\hat{\beta}_{is_i}^m = \delta_{is_i} \equiv 0$ for $i = 1, \cdots, w$.

Consider the WM-based decision function $\hat{h}^m(\boldsymbol{x}) = \tilde{\beta}_0 + \boldsymbol{x}\hat{\boldsymbol{\beta}}^m = h(\boldsymbol{x}) + \boldsymbol{x}\Delta\hat{\boldsymbol{\beta}}^m$. The above derivation shows that $\sqrt{n}(\boldsymbol{x}\Delta\hat{\boldsymbol{\beta}}^m - \boldsymbol{x}\boldsymbol{\delta}) \xrightarrow{L} \mathcal{N}(0, V^f(\hat{\boldsymbol{\beta}}^m, \boldsymbol{x}))$, where the variance is determined by the covariance matrix of $\Delta\hat{\boldsymbol{\beta}}^m$. Following Theorem 2, the expectation of the total incremental error rate of the WM procedure

$$\mathbb{E}[\Delta R^f(\hat{\boldsymbol{\beta}}^m)] \to \sum_{\mathcal{B}^f(\hat{\boldsymbol{\beta}}^m)} \tanh|h(\boldsymbol{x})/2|\ \boldsymbol{\theta}_0(\boldsymbol{x}), \text{ as } n \to \infty, \tag{33}$$

where $\mathcal{B}^f(\hat{\boldsymbol{\beta}}^m) = \{\boldsymbol{x} : \text{sign}\{h(\boldsymbol{x})\}\{h(\boldsymbol{x}) + \boldsymbol{x}\boldsymbol{\delta}\} < 0\}$. Similarly, the incremental percentage over the ideal error rate $\mathbb{E}[\Delta R^f(\hat{\boldsymbol{\beta}}^m)]/(R^f)^*$ is used to compare the predictions of the WM and the FE procedures, given that $\mathbb{E}[\Delta R^f(\hat{\boldsymbol{\beta}}^f)] \to 0$ (28). Under the FEM, the ideal error rate

$$(R^f)^* = \mathbb{E}_{\boldsymbol{\theta}_0}\left\{p_0(\boldsymbol{X})\mathbf{1}(h(\boldsymbol{X}) \ge 0) + p_1(\boldsymbol{X})\mathbf{1}(h(\boldsymbol{X}) < 0)\right\}. \tag{34}$$

Recall that $p_y(\boldsymbol{X}) = P(Y = y \mid \boldsymbol{X})$, i.e.,

$$p_y(\boldsymbol{X}) = \frac{e^{yh(\boldsymbol{X})}}{e^{h(\boldsymbol{X})} + 1}, \text{ for } y = 0, 1.$$

3.4. *FEM with Markov background.* Next, we generalize the FEM to Markov background and assume that any sequence $\boldsymbol{X} \in \mathcal{I}^w$ is generated marginally by a Markov chain. Consistent with Section 3.2, we denote by $\boldsymbol{\psi}_0 = (\psi_0(x,y))_{J\times J}$ the transition probability matrix of the Markov chain.

It is trivial to see that, with the Markov background model, the ideal decision is still $h(\boldsymbol{x}) = \tilde{\beta}_0 + \boldsymbol{x}\tilde{\boldsymbol{\beta}}$. If $V_1^f(\hat{\boldsymbol{\beta}}^f, \boldsymbol{x})$ denotes the asymptotic variance of $\sqrt{n}\boldsymbol{x}d\hat{\boldsymbol{\beta}}^f$ under Markov background, then with V_1^f in place of V^f equation (28) remains valid for the FE-based prediction. On the other hand, if we proceed with the WM procedure, the expected incremental error rate

$$\mathbb{E}[\Delta R_1^f(\hat{\boldsymbol{\beta}}^m)] \to \sum_{\mathcal{B}_1^f(\hat{\boldsymbol{\beta}}^m)} \tanh|h(\boldsymbol{x})/2|\ \boldsymbol{\psi}_0(\boldsymbol{x}),\ \text{as } n\to\infty, \tag{35}$$

where $\mathcal{B}_1^f(\hat{\boldsymbol{\beta}}^m) = \{\boldsymbol{x} : \text{sign}\{h(\boldsymbol{x})\}\{h(\boldsymbol{x})+\delta(\boldsymbol{x})\} < 0\}$. Here $\delta(\boldsymbol{x})$ denotes the asymptotic bias of the WM-based decision function for $\boldsymbol{x}$. A detailed derivation of equation (35) and the bias $\delta(\boldsymbol{x})$ is provided in Appendix B. In analogy to the FEM with i.i.d. background, $\mathbb{E}[\Delta R_1^f(\hat{\boldsymbol{\beta}}^m)]/(R_1^f)^*$ measures the increased error rate of the WM procedure relative to the FE procedure, where

$$(R_1^f)^* = \mathbb{E}_{\boldsymbol{\psi}_0}\left\{p_0(\boldsymbol{X})\mathbf{1}(h(\boldsymbol{X}) \ge 0) + p_1(\boldsymbol{X})\mathbf{1}(h(\boldsymbol{X}) < 0)\right\}.$$

3.5. *Mis-specification of the constant term.* In all the above derivations, we have assumed that the constant term $\beta_0(\tilde{\beta}_0)$ is fixed to its true value. If this is not the case, then the deviation $\Delta\hat{\beta}_0 = \hat{\beta}_0 - \beta_0(\tilde{\beta}_0)$ will be an extra bias term for an estimated decision in which the constant term is fixed to $\hat{\beta}_0$. More specifically, the set $\mathcal{B}^f(\hat{\boldsymbol{\beta}}^m)$ in equation (33) will be replaced by $\{\boldsymbol{x} : \text{sign}\{h(\boldsymbol{x})\}\{h(\boldsymbol{x}) + \boldsymbol{x}\boldsymbol{\delta} + \Delta\hat{\beta}_0\} < 0\}$, and similarly for $\mathcal{B}_1^m(\hat{\boldsymbol{\beta}}^f)$ in equation (26) and $\mathcal{B}_1^f(\hat{\boldsymbol{\beta}}^m)$ in equation (35).

4. Computation. To apply the theoretical results, we need to solve the minimization (19) and the summation (20) involved in Theorems 1 and 2, respectively. If the width of a motif $w \le 12$, brute-force enumeration of all w-mers is computationally feasible, which provides exact solutions for both the minimization and the summation problems.

For a motif of width $w > 12$, we minimize (19) to find $\alpha(\hat{h})$ by a two-step approach. We generate $N = 5\times 10^6$ w-mers from the background model $\boldsymbol{\theta}_0$ and identify the minimum of (19) among them. Then we refine the obtained minimum by simulated annealing for 5,000 iterations with temperature decreasing linearly from one to zero. At each iteration, we randomly choose one nucleotide X_i from the w positions and propose to mutate X_i to one of the other three nucleotide bases with equal probability. The proposal is accepted according to a Metropolis-Hastings ratio with current temperature.

Since the set $\mathcal{B}(\hat{h})$, as in equations (26), (33) and (35), is usually small, it will be very inefficient to approximate the summation by generating w-mers from background distributions. Thus, we develop an importance sampling approach to approximate the summation (20) when $w > 12$. Here, we use the calculation of $\mathbb{E}[\Delta R^f(\hat{\boldsymbol{\beta}}^m)]$ (33) to

illustrate this approach. Note that one can bound $\boldsymbol{x\delta}$ in the definition of $\mathcal{B}^f(\hat{\boldsymbol{\beta}}^m)$ so that $\boldsymbol{x\delta} \in [M_1, M_2]$, where $M_1 = \sum_{i=1}^w \min_j \delta_{ij}$ and $M_2 = \sum_{i=1}^w \max_j \delta_{ij}$. These bounds imply that if $\boldsymbol{x} \in \mathcal{B}^f(\hat{\boldsymbol{\beta}}^m)$ then

$$h(\boldsymbol{x}) \in (-M_2, 0) \cup (0, -M_1) := \mathcal{H}.$$

We design a sequential proposal $g(\boldsymbol{X})$ that is more likely to generate $\boldsymbol{X}$ with $h(\boldsymbol{X}) \in \mathcal{H}$. Suppose that we have generated $X_1, \cdots, X_{k-1}$ $(1 \le k \le w)$ from this proposal. Let $h_{k-1} = \tilde{\beta}_0 + \sum_{i=1}^{k-1} \tilde{\beta}_{iX_i}$, in particular $h_0 = \tilde{\beta}_0$. We determine $B_{k+1}^{(L)} = \sum_{i>k} \min_j \tilde{\beta}_{ij}$ and $B_{k+1}^{(U)} = \sum_{i>k} \max_j \tilde{\beta}_{ij}$, the bounds for $\sum_{i>k} \tilde{\beta}_{iX_i}$. If $X_k = j$ then the range for $h(\boldsymbol{X})$ is

$$h_{k-1} + \tilde{\beta}_{ij} + [B_{k+1}^{(L)}, B_{k+1}^{(U)}] := [L_{kj}, U_{kj}].$$

The larger the overlap between this interval and $\mathcal{H}$, the more likely that $\boldsymbol{X}$ will belong to the desired set $\mathcal{B}^f(\hat{\boldsymbol{\beta}}^m)$. Thus, we propose X_k with probability

$$g_k(X_k = j \mid X_1, \cdots, X_{k-1}) \propto |[L_{kj} - \epsilon, U_{kj} + \epsilon] \cap \mathcal{H}|, \tag{36}$$

where $|\cdot|$ returns the length of an interval and ϵ is a small positive number to allow the generation of $X_k = j$ when $L_{kj} = U_{kj} \in \mathcal{H}$. Proposing X_k sequentially by (36) for $k = 1, \cdots, w$ generates an $\boldsymbol{X}$ from $g(\boldsymbol{X})$. With N proposed samples $\{\boldsymbol{X}^{(t)}\}_{t=1}^N$ we estimate the summation (33) by

$$\frac{1}{N} \sum_{t=1}^{N} \tanh \left| h(\boldsymbol{X}^{(t)})/2 \right| \frac{\boldsymbol{\theta}_0(\boldsymbol{X}^{(t)}) \mathbf{1}\{\boldsymbol{X}^{(t)} \in \mathcal{B}^f(\hat{\boldsymbol{\beta}}^m)\}}{g(\boldsymbol{X}^{(t)})}.$$

In this work, we propose $N = 5 \times 10^6$ samples for this importance sampling estimation. We verified that the estimations were very close to the exact summations. With different bounds for $h_1(\boldsymbol{x})$ and $h(\boldsymbol{x})$, this approach is applied to other similar summations in (26) and (35).

5. Numerical study. A numerical study was performed under the WMM to confirm and quantify the lower predictive efficiency of the FE-based estimator $\hat{\boldsymbol{\beta}}^f$ compared to the WM-based estimator $\hat{\boldsymbol{\beta}}^m$ discussed in Section 3.1. We randomly selected 200 TFs from the database TRANSFAC (Matys et al. 2003). For each TF, experimentally verified binding sites were used to construct a weight matrix with a small amount of pseudo counts. Then we randomly sampled 5,000 human upstream sequences, each of length 10 kilo bases, and calculated their nucleotide frequency $\hat{\boldsymbol{\theta}}_0 =$ (0.263, 0.234, 0.237, 0.266). The 200 weight matrices display large variability. The width w ranges from 6 to 21 and the information content, $\sum_{i=1}^w \{2 + \mathbb{E}_{\boldsymbol{\theta}_i}(\log_2 \theta_{iX_i})\}$, ranges from 5.1 to 17.5 bits (Figure 1). These statistics show that our selection has covered the typical width and strength of DNA motifs.

A constructed weight matrix was regarded as the parameter $\boldsymbol{\Theta}$ and the nucleotide frequency $\hat{\boldsymbol{\theta}}_0$ was used for the i.i.d. background in the WMM. Since the prior odds ratio (q_1/q_0) of a binding site over a background site is usually small, we chose three typical

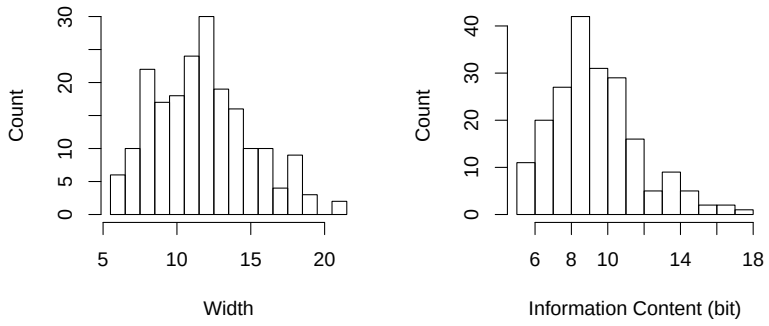

FIG 1. *Histograms of the width and the information content of 200 WMs.*

TABLE 2
Summary of $ARE^m(\hat{\boldsymbol{\beta}}^f, \hat{\boldsymbol{\beta}}^m)$

λ	Min.	Q_1	Median	Q_3	Max.
200	0.134	0.391	0.490	0.595	0.849
500	0.183	0.475	0.555	0.638	0.849
1000	0.217	0.508	0.560	0.675	0.918

$Q_{1,3}$: the first and the third quartiles.

values for the inverse of the prior odds, $\lambda = q_0/q_1 = 200, 500, 1000$, for numerical calculations. We evaluated the AREs of the FE-based prediction w.r.t. the WM-based prediction, defined by $\mathrm{ARE}^m(\hat{\boldsymbol{\beta}}^f, \hat{\boldsymbol{\beta}}^m) = \alpha^m(\hat{\boldsymbol{\beta}}^f)/\alpha^m(\hat{\boldsymbol{\beta}}^m)$ in Section 3.1, for the 200 WMs. As discussed in Section 4, our evaluation of AREs was exact for WMs of $w \leq 12$ and was carried out with simulated annealing for $w > 12$. In addition, Monte Carlo average was utilized, before simulated annealing, to approximate $\mathrm{Cov}^m(\sqrt{n}d\hat{\boldsymbol{\beta}}^f)$ (22) by simulating 5×10^6 w-mers from the i.i.d. background.

The asymptotic relative efficiencies $\mathrm{ARE}^m(\hat{\boldsymbol{\beta}}^f, \hat{\boldsymbol{\beta}}^m)$ on the 200 TFs are summarized in Table 2 for the three inverse prior odds. It is seen that for all the WMs the FE-based prediction shows lower efficiency than the WM-based prediction, and that the median AREs of $\hat{\boldsymbol{\beta}}^f$ to $\hat{\boldsymbol{\beta}}^m$ are between 50% and 60% and the third quartiles (Q_3) between 60% and 70%. Thus, for more than 75% of the TFs, the FE procedure is less than 70% as efficient as the WM procedure in terms of prediction. This confirms the loss of efficiency of the FE-based prediction under the WMM, although both estimators are consistent. We note that the increase of ARE with higher λ (smaller q_1) is consistent with the lower bound defined in Proposition 3.

6. Applications. In this section, we apply the WM and the FE approaches to ChIP-seq data and protein binding microarray (PBM) data. We perform cross validation (CV) with training data of different size, ranging from 20 to 500 binding sites, for two purposes. First, with the large scale of both types of data, we can compare

empirical error rates in cross validation against theoretical error rates. This may allow us to verify some of the model assumptions and propose further improvement on the models. Second, we are also interested in examining the practical performance of the two computational methods when the number of observed binding sites varies in a wide range, which will provide useful guidance for future applications.

6.1. *ChIP-seq data.* In the recent two years, the ChIP-seq technique (Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007) has become a powerful high-throughput method to detect TFBS's in whole genome scale. A binding peak in ChIP-seq data can usually narrow down the location of a TFBS to a neighborhood of 50 to 200 bps (Johnson et al., 2007). ChIP-seq data that contain thousands of binding sites for a number of TFs have been generated in a study on mouse embryonic stem cells (Chen et al., 2008). We chose five TFs, Esrrb, Oct4, STAT3, Sox2 and cMyc, in this study to compare the WM and the FE methods. The five TFs all have well-defined weight matrices in literature and each contains more than 2,000 detected binding peaks in ChIP-seq, and their data quality was confirmed by motif enrichment analysis in Chen et al. (2008). To identify the exact binding site of a ChIP-seq binding peak, we searched the 200-bp neighborhood of the peak, 100 bps on each side, to find the best match to the known weight matrix of the TF. Given the very small search space, the uncertainty in the exact location of the binding site should be minimal. If the motif width of a TF is w, background w-mers were extracted from genomic control regions that match the locations of the binding sites relative to nearby genes. The ratio of the number of background sites over the number of binding sites was set to 200 for every TF, that is, the inverse prior odds ratio $\lambda = q_0/q_1 = 200$. A transition matrix was estimated from the extracted background sites for each TF, since the log Bayes factor of a Markov background model over an i.i.d model was $> 10^5$.

Based on the way we composed the data sets, the WMM with Markov background (Section 3.2) seems a more plausible data generation model. Clearly, a data set was a mixture of detected binding sites and random background sites, and the background distribution was close to a Markov chain. If there is no within-motif dependence, binding sites can be regarded as being generated from a WM model, and consequently, the WM-based prediction is expected to have a smaller error rate compared to the FE-based prediction. However, if there exists within-motif dependence in binding sites, the FEM, which is able to capture such dependence, may outperform the WM approach regardless of the mixture nature of the data sets. We computed theoretical error rates of the two approaches under the WMM with Markov background. For each TF, we estimated a WM from all the binding sites and a transition matrix from the background sites. Regarding them as the model parameters, we calculated the asymptotic error rate of the WM-based prediction, which is the ideal error rate (27), and the incremental rate of the FE-based prediction (26). Note that the bias due to mis-specification of the constant term in the FE approach needs to be included for the calculation of equation (26). These theoretical error rates are reported in Table 3 (the column of $n^+ = \infty$).

To compare with theoretical results, we performed cross validation to compute empirical error rates of the WM and the FE procedures on each data set. We randomly

TABLE 3
Predictive error rates (in the unit of 10^{-3}) for ChIP-seq data

TF	n^+	20	50	100	200	500	∞
Esrrb	WM	2.66	2.49	2.43	2.40	2.36	2.30
	FE	4.30	2.77	2.54	2.45	2.37	2.45
	(FE-WM)/WM (%)	61.7	11.2	4.5	2.1	0.4	6.5
Oct4	WM	3.68	3.54	3.50	3.47	3.44	2.98
	FE	4.81	3.71	3.53	3.45	3.41	3.06
	(FE-WM)/WM (%)	30.7	4.8	0.9	−0.6	−0.9	2.7
STAT3	WM	3.03	2.84	2.78	2.72	2.69	2.57
	FE	4.69	3.09	2.84	2.75	2.70	2.74
	(FE-WM)/WM (%)	54.8	8.8	2.2	1.1	0.3	6.6
Sox2	WM	2.95	2.75	2.68	2.65	2.63	2.53
	FE	3.44	2.89	2.73	2.68	2.66	2.59
	(FE-WM)/WM (%)	16.6	5.1	1.9	1.1	1.1	2.4
cMyc	WM	2.67	2.49	2.42	2.38	2.34	2.07
	FE	3.30	2.51	2.34	2.26	2.23	2.24
	(FE-WM)/WM (%)	23.6	0.8	−3.3	−5.0	−4.7	8.2

Note: Reported are average error rates over 100 CVs.

sampled (without replacement) n^+ binding sites and $\lambda \cdot n^+$ background sites from a full data set to form a training set. Both approaches were applied to the training set to estimate their respective decision functions. For WM-based prediction, a WM and a transition matrix were estimated from the training data set to construct a decision function (24) with $\beta_0 = -\log(\lambda)$. For FE-based prediction, we applied logistic regression to the training set to obtain $\hat{h}^f(\boldsymbol{x}) = \hat{\beta}_0 + \boldsymbol{x}\hat{\boldsymbol{\beta}}^f$. Then we predicted the class labels of the remaining unused sequences (test set) by each of the two decision functions and calculated empirical error rates (CV error rates). This procedure was repeated 100 times independently for each value of n^+ to obtain the average CV error rate. To examine performance with a varying sample size (the number of sequences in a training set), we chose n^+ from 20 to 500.

The average CV error rates are reported in Table 3. The theoretical results give a reasonable approximation to the CV error rates for both approaches when the training sample size $n^+ \geq 200$. The asymptotic error rates of the WM approach are uniformly lower than its CV error rates for all the TFs, while the FE approach achieves a smaller CV error rate with $n^+ = 500$ than its asymptotic rate for three TFs. Consequently, the incremental percentage of the FE-based prediction for $n^+ = 500$ is less than the expected level calculated from the theory. This comparison implies that the WMM may not match the exact underlying data generation process, although it is more plausible than the FEM given the mixture composition of the data sets. As we discussed, potential dependence within a motif may cause possible violation to the WMM. To verify our hypothesis, we conducted the χ^2-test for every pair of motif positions (X_i and X_k, $1 \leq i < k \leq w$) given the binding sites in each data set. At the significance level of 0.005, we identified 25, 19, 17, 8, and 12 pairwise correlations for Esrrb, Oct4, STAT3, Sox2, and cMyc binding sites, respectively, which gives a false discovery rate of $< 2\%$

for all the TFs. By capturing such correlations the FEM is able to achieve comparable or even slightly better prediction than the WMM with a moderate-size training sample ($n^+ \geq 100$, Table 3). Finally, it is important to note that even under the exact model assumptions of the WMM, the FE-based prediction only results in a marginal increment in error rate ($< 10\%$) compared to the WM approach asymptotically (Table 3, $n^+ = \infty$). Together with the superior or comparable CV performance when the training size is reasonably large, this result suggests the use of the FE approach, when we have a sufficient number of observed binding sites.

6.2. *PBM data.* Protein binding microarrays (Mukherjee et al., 2004) provide a high throughput means to interrogate protein binding specificity to DNA sequences. Quantitative measurement of the binding specificity of a protein to every short nucleotide sequence designed on a DNA microarray can be obtained simultaneously. The PBM data in Berger et al. (2008) quantified DNA binding of homeodomain proteins via the calculation of an enrichment score, with an expected false discovery rate (FDR), for each double-stranded nucleotide sequence of length eight ($w = 8$). The data set for each protein contains 32,896 8-mers, each with an enrichment score and an FDR. We identified as the consensus binding pattern for a protein the 8-mer with the highest enrichment score, and then labeled as binding sites those 8-mers whose FDR < 0.005 and which differ by no more than three nucleotides from the consensus after considering both the forward and the reverse complement strands. The remaining 8-mers were labeled as background sites and we randomly determined their strands (orientations) to avoid potential artifacts. In this study we included five proteins, Hoxa11, Irx3, Lhx3, Nkx2.5, and Pou2f2, each from a different family, and called 134, 190, 267, 145, and 213 binding sites, respectively.

The FEM, developed by the biophysics of protein-DNA binding, is expected to be a better model that matches the design of PBM data than the WMM. Thus, theoretical analysis was conducted under the FEM for the five PBM data sets. We applied logistic regression to estimate $\tilde{\boldsymbol{\beta}}$ and $\tilde{\beta}_0$ (9) with all the labeled 8-mers in a data set, where the 8-mer 'AAAAAAAA' was regarded as the reference sequence, i.e., $\beta_{i1} \equiv 0$. We calculated the ideal error rate $(R^f)^*$ (34) of the FE-based prediction, with an i.i.d. uniform background (by design the background distribution is uniform). For the WM approach, we chose $\hat{\beta}_0$ as the log-ratio of the number of binding sites over that of background sites, and calculated its asymptotic error rate by equation (33), in which the bias in the constant term ($\Delta\hat{\beta}_0$) was included. The theoretical error rates are reported in Table 4 ($n^+ = \infty$), where we find that the WM approach gives a significantly higher error rate, between 14% and 56%, than the FE approach.

The same CV procedure as in the previous section was performed on the PBM data sets to compare the empirical predictive error rates of the two approaches, with n^+ varying between 20 and 100 (Table 4). There is a clear decreasing trend in error rate for both approaches with the increase of the training sample size n^+, although for some data sets the difference between the CV error rate for $n^+ = 100$ and the asymptotic rate is still quite obvious. Such discrepancy is probably due to the following two reasons. First, the parameters $(\tilde{\boldsymbol{\beta}}, \tilde{\beta}_0)$ used for the calculation of asymptotic rates were

TABLE 4
Predictive error rates (in the unit of 10^{-3}) for PBM data

Protein	n^+	20	50	100	∞
Hoxa11	FE	3.57	2.62	2.43	1.63
	WM	3.51	3.22	3.05	2.10
	(WM-FE)/FE (%)	−1.7	22.9	25.5	28.8
Irx3	FE	5.91	4.71	4.54	3.26
	WM	5.06	4.85	4.79	3.71
	(WM-FE)/FE (%)	−14.4	3.0	5.5	13.8
Lhx3	FE	7.51	4.64	4.20	3.18
	WM	6.90	6.54	6.47	4.97
	(WM-FE)/FE (%)	−8.1	40.9	54.0	56.3
Nkx2.5	FE	3.73	2.31	2.12	1.80
	WM	3.64	3.29	3.16	2.36
	(WM-FE)/FE (%)	−2.4	42.4	49.1	31.1
Pou2f2	FE	6.25	4.91	4.56	3.31
	WM	5.79	5.57	5.49	4.02
	(WM-FE)/FE (%)	−7.3	13.4	20.4	21.5

estimated from data sets which only contain 100 to 200 binding sites. This resulted in a high variance in the estimated parameters: The median ratio of the standard error over the absolute value of an estimated coefficient was between 10% and 30% for the five data sets. Second, the training sample size, $n^+ = 100$, is still too small to achieve a comparable error rate as $n^+ \to \infty$. However, we have already seen substantially increased error rates of the WM-based predictions compared to the FE-based predictions for $n^+ = 100$, which is very consistent with the theoretical results. This comparison confirms that unless the training sample size is really small, using the WM approach may degrade predictive performance dramatically if the data generation mechanism is close to the FEM.

7. Discussion. Combining results on the ChIP-seq data and the PBM data, this study provides some general guidance for practical applications of the WM and the FE approaches, irrespective of underlying data generation. When the training sample size is small, the WM procedure seems to produce fewer errors than the FE procedure. But when we have observed enough binding sites, the advantage of the FE procedure is clearly seen. On one hand, it gives a comparable or slightly better prediction than the WM approach even if the WMM is more likely for the data (Table 3, $n^+ \geq 100$). On the other hand, when the data are generated in a way that matches the biophysical process of protein-DNA binding such as the PBM data, the reduction in error rate of the FE approach can be substantial compared to the WM approach (Table 4, $n^+ \geq 50$). The relative performance between the two approaches reflects a typical variance-bias tradeoff. Estimation under the WMM is simple and more robust, which typically has a smaller variance than the FEM. For a small sample size, predictive errors are mostly caused by variance in estimation and thus, WM-based predictions may outperform FE-based predictions. When the sample size increases, estimation variance decreases for both approaches and the potential bias in the WM approach

becomes the main factor for predictive errors. Given that its primary principle comes from the biophysics of protein-DNA interactions, the FEM has become more attractive, based on which many computational methods have been developed for predicting TF-DNA binding. In these methods a weight matrix is sometimes used as a first order approximation for computing free energy-based binding affinity. This work suggests that this approximation must be applied with caution. The results on the PBM data have demonstrated that the WM procedure may give a prediction with 50% or more errors compared to the FE-based decision for a reasonably large sample size (Table 4).

In recent years, a substantial amount of large-scale TF-DNA binding data have been generated for many important biological processes. As demonstrated by the applications to ChIP-seq data and PBM data, large-sample theory is able to provide valuable insights on statistical estimation and prediction for such large-scale data. The results in this article can be regarded as a first step towards a theoretical development on computational approaches for gene regulation analysis. Incorporation of within-motif dependence in the WMM and interaction effects in the FEM is a direct next step of this work, for which the model selection component needs to be considered in a theoretical analysis. Although desired, further generalizations to methods for *de novo* motif discovery, identification of *cis*-regulatory modules and predictive modeling of gene regulation will be more challenging future directions.

Appendices.

Appendix A: Proof of Proposition 3.

PROOF. Let $\theta_{k(-j)} = 1 - \theta_{kj}$ for $k = 0, i$. The second order partial derivative of the marginal log-likelihood $l(\boldsymbol{\Theta} \mid \boldsymbol{X}) = \log\{q_0\boldsymbol{\theta}_0(\boldsymbol{X}) + q_1\boldsymbol{\Theta}(\boldsymbol{X})\}$ w.r.t. θ_{ij} is

$$\frac{\partial^2 l(\boldsymbol{\Theta} \mid \boldsymbol{X})}{\partial \theta_{ij}^2} = -\frac{q_1^2\{\boldsymbol{\Theta}_{[-i]}(\boldsymbol{X}_{[-i]})\}^2}{\{q_0\boldsymbol{\theta}_0(\boldsymbol{X}) + q_1\boldsymbol{\Theta}(\boldsymbol{X})\}^2},$$

where $\boldsymbol{\Theta}(\boldsymbol{X}) = \boldsymbol{\Theta}_{[-i]}(\boldsymbol{X}_{[-i]}) \cdot \theta_{iX_i}$ for $X_i = j, (-j)$ and similarly for $\boldsymbol{\theta}_0(\boldsymbol{X})$. Thus, the Fisher information on θ_{ij} given $\boldsymbol{X}$ is

$$\begin{aligned} I(\theta_{ij} \mid \boldsymbol{X}) &= -\mathbb{E}\left\{\frac{\partial^2 l(\boldsymbol{\Theta} \mid \boldsymbol{X})}{\partial \theta_{ij}^2}\right\} = \sum_{\boldsymbol{x}} \frac{q_1^2\{\boldsymbol{\Theta}_{[-i]}(\boldsymbol{x}_{[-i]})\}^2}{q_0\boldsymbol{\theta}_0(\boldsymbol{x}) + q_1\boldsymbol{\Theta}(\boldsymbol{x})} \\ &= q_1 \sum_{\boldsymbol{x}} \frac{q_1\boldsymbol{\Theta}(\boldsymbol{x})}{q_0\boldsymbol{\theta}_0(\boldsymbol{x}) + q_1\boldsymbol{\Theta}(\boldsymbol{x})} \cdot \frac{1}{\theta_{ix_i}} \cdot \boldsymbol{\Theta}_{[-i]}(\boldsymbol{x}_{[-i]}) \\ &= q_1 \sum_{x \in \{j,(-j)\}} \frac{1}{\theta_{ix}} \cdot \mathbb{E}_{\boldsymbol{\Theta}_{[-i]}}\left\{\left(\frac{q_0\theta_{0x}\boldsymbol{\theta}_0(\boldsymbol{X}_{[-i]})}{q_1\theta_{ix}\boldsymbol{\Theta}_{[-i]}(\boldsymbol{X}_{[-i]})} + 1\right)^{-1}\right\}. \end{aligned}$$

Because $\mathbb{E}_{\boldsymbol{\Theta}_{[-i]}}\left\{\boldsymbol{\theta}_0(\boldsymbol{X}_{[-i]})/\boldsymbol{\Theta}_{[-i]}(\boldsymbol{X}_{[-i]})\right\} = 1$, Jensen's inequality implies that

$$I(\theta_{ij} \mid \boldsymbol{X}) \geq \sum_{x \in \{j,(-j)\}} \frac{q_1^2}{q_0\theta_{0x} + q_1\theta_{ix}} = \frac{q_1^2}{\bar{\theta}_{ij}(1 - \bar{\theta}_{ij})}.$$

The lower bound $B(q_1, \theta_{ij}, \theta_{0j})$ is obtained by dividing the R.H.S. of this inequality by the Fisher information on θ_{ij} given $\boldsymbol{X}$ and Y jointly,

$$I(\theta_{ij} \mid \boldsymbol{X}, Y) = -\mathbb{E}\left\{\frac{\partial^2 l(\boldsymbol{\Theta} \mid \boldsymbol{X}, Y)}{\partial \theta_{ij}^2}\right\} = \frac{q_1}{\theta_{ij}(1-\theta_{ij})},$$

where $l(\boldsymbol{\Theta} \mid \boldsymbol{X}, Y) = \log P(\boldsymbol{X}, Y \mid \boldsymbol{\Theta})$ is the joint log-likelihood. □

Appendix B: Derivation of $\mathbb{E}[\Delta R_1^f(\hat{\boldsymbol{\beta}}^m)]$ *(35).* Given the estimated weight matrix $\hat{\boldsymbol{\Theta}}^m$ based on observed binding sites $\boldsymbol{D}_n^+$, the constructed decision function of the WM approach

$$\begin{aligned}\hat{h}_1^m(\boldsymbol{x}) &= \log(q_1/q_0) + \sum_{i=1}^{w}[\log \hat{\theta}_{ix_i}^m - \log \psi_0(x_{i-1}, x_i)] \\ &\xrightarrow{P} \tilde{\beta}_0 + \sum_{i=1}^{w}\left\{\log \frac{\theta_{ix_i}^f}{\theta_{is_i}^f} - \log \frac{\psi_0(x_{i-1}, x_i)}{\psi_0(s_{i-1}, s_i)}\right\}, \text{ as } n \to \infty, \end{aligned} \tag{37}$$

where $\theta_{ij}^f = P(X_i = j \mid Y = 1)$ under the FEM with Markov background, $(s_1, \cdots, s_w)$ is the reference sequence, and

$$\tilde{\beta}_0 = \log(q_1/q_0) + \sum_{i=1}^{w} \log\{\theta_{is_i}^f / \psi_0(s_{i-1}, s_i)\}.$$

Let $\boldsymbol{x}_{[-i]}(s) = (x_1, \cdots, x_{i-1}, s, x_{i+1}, \cdots, x_w)$. Following a similar derivation in Section 3.3, we have $\theta_{ij}^f \propto \exp(\tilde{\beta}_{ij} + \eta_{ij})$, where

$$\eta_{ij} = \log\left\{\sum_{\boldsymbol{x}_{[-i]}} \frac{e^{u_i}}{1 + e^{\tilde{\beta}_{ij}} e^{u_i}} \boldsymbol{\psi}_0(\boldsymbol{x}_{[-i]}(j))\right\} - \log\left\{\sum_{\boldsymbol{x}_{[-i]}} \frac{e^{u_i}}{1 + e^{u_i}} \boldsymbol{\psi}_0(\boldsymbol{x}_{[-i]}(s_i))\right\}$$

with $u_i = \tilde{\beta}_0 + \boldsymbol{x}_{[-i]}\tilde{\boldsymbol{\beta}}_{[-i]}$. Since $\tilde{\beta}_{is_i} = \eta_{is_i} = 0$, $\log(\theta_{ij}^f/\theta_{is_i}^f) = \tilde{\beta}_{ij} + \eta_{ij}$ for all i and j. Thus, equation (37) becomes

$$\hat{h}_1^m(\boldsymbol{x}) \xrightarrow{P} \tilde{\beta}_0 + \sum_{i=1}^{w}\left\{\tilde{\beta}_{ix_i} + \eta_{ix_i} - \log \frac{\psi_0(x_{i-1}, x_i)}{\psi_0(s_{i-1}, s_i)}\right\} = h(\boldsymbol{x}) + \delta(\boldsymbol{x}),$$

where $\delta(\boldsymbol{x}) = \sum_{i=1}^{w} \eta_{ix_i} - \log\{\psi_0(x_{i-1}, x_i)/\psi_0(s_{i-1}, s_i)\}$. Let $\Delta\hat{h}_1^m(\boldsymbol{x}) = \hat{h}_1^m(\boldsymbol{x}) - h(\boldsymbol{x})$. The asymptotic normality of $\sqrt{n}d\hat{\boldsymbol{\Theta}}^m$ implies that $\sqrt{n}\{\Delta\hat{h}_1^m(\boldsymbol{x}) - \delta(\boldsymbol{x})\}$ follows a limiting normal distribution with mean 0 and a finite (possibly zero) variance for every $\boldsymbol{x}$. Equation (35) then follows from Theorem 2.

Acknowledgements. The author thanks Wing H. Wong, Jun S. Liu and Zhengqing Ouyang for helpful discussions. This work was supported by NSF grant DMS-0805491.

References.

[1] Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of 2nd International Conference on Intelligent Systems for Molecular Biology*, 28-36. CA: AAAI Press.

[2] Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependence in protein-DNA binding sites. *RECOMB 2003*, Berlin, Germany.

[3] Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Research*, 30, 442-451.

[4] Berg, O.G. and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, 193, 723-750.

[5] Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A. et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133, 1266-1276.

[6] Bulyk, M.L., Johnson, P.L.F., and Church, G.M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, 30, 1255-1261.

[7] Bussemaker, H.J., Foat, B.C., and Ward, L.D. (2007). Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annual Review of Biophysics and Biomolecular Structure*, 36, 329-347.

[8] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B. et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133, 1106-1117.

[9] Djordjevic, M., Sengupta, A.M., and Shraiman, B.I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Research*, 13, 2381-2390.

[10] Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70, 892-898.

[11] Elnitski, L., Jin, V.X., Farnham, P.J., and Jones, S.J.M. (2006). Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Research*, 16, 1455-1464.

[12] Ferguson, T.S. (1996). *A Course in Large Sample Theory*, p. 105-139. London: Chapman & Hall.

[13] Foat, B.C., Morozov, A. and Bussemaker, H.J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22, e141-e149.

[14] Gerland, U., Moroz, J.D., and Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proceedings of the National Academy of Sciences USA*, 99, 12015-12020

[15] Granek, J.A. and Clarke, N.D. (2005). Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biology*, 6, R87.

[16] Hertz, G.Z. and Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15, 563-577.

[17] Ji, H.K. and Wong, W.H. (2006). Computational biology: toward deciphering gene regulatory information in mammalian genomes. *Biometrics*, 62, 645-663.

[18] Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316, 1497-1502.

[19] Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31, 3576-3579.

[20] Kinney, J.B., Tkacik, G., Callan, C.G. Jr (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences USA*, 104, 501-506.

[21] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wooton, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262, 208-214.

[22] Liu, X.S., Brutlag, D.L., and Liu, J.S. (2002). An algorithm for finding protein-DNA binding sites

with applications to chromatin immunoprecipitation microarray experiments. *Nature Biotechnology*, 20, 835-839.

[23] Matys, V., Fricke, E., Geffers, R., Göβling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. et al. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31, 374-378.

[24] Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G. et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-comitted cells. *Nature*, 448, 553-559.

[25] Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M. et al. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics*, 36, 1331-1339.

[26] Rahmann, S., Muller, T., and Vingron, M. (2003). On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology*, 2, Article 7.

[27] Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T. et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massive parallel sequencing. *Nature Methods*, 4, 651-657.

[28] Roider, H.G., Manke, T., and Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23, 134-141.

[29] Roth, F.R., Hughes, J.D., Estep, P.E., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole genome mRNA quantization. *Nature Biotechnology*, 16, 939-945.

[30] Stormo, G.D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, 16, 16-23.

[31] Stormo, G.D. and Fields, D.S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends in Biochemical Sciences*, 23, 109-113.

[32] Stormo, G.D. and Hartzell, G.W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences USA*, 86, 1183-1187.

[33] Turatsinze, J.V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, 3, 1578-1588.

[34] Vingron, M., Brazma, A., Coulson, R., van Helden, J., Manke, T., Palin, K., Sand, O., and Ukkonen, E. (2009). Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biology*, 10, 202.

[35] von Hippel, P.H. and Berg, O.G. (1986). On the specificity of DNA-protein interactions. *Proceedings of the National Academy of Sciences USA*, 83, 1608-1612.

[36] Zhao, X., Huang, H., and Speed, T.P. (2005). Finding short DNA motifs using permuted Markov models. *Journal of Computational Biology*, 12, 894-906.

[37] Zhou, Q. and Liu, J.S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20, 909-916.

[38] Zhou, Q. and Liu, J.S. (2008). Extracting sequence features to predict protein-DNA interactions: a comparative study. *Nucleic Acids Research*, 36, 4137-4148.